Parameter Estimation for Self-Assembly via Simulation-Based Data Fitting Methods

Lu Xie[1,3], Gregory R. Smith[2], and Russell Schwartz[2, 3]

***Abstract* — Self-assembly is a pervasive phenomenon in life sciences but difficult to model due to the enormous number of possible reaction pathways they may use and our limited ability to distinguish them experimentally. Simulation methods offer great help for revealing possible assembly pathways, but depend on physical parameters that one cannot directly measure. We address the problem of fitting kinetic rate parameters for rule-based stochastic simulations to indirect bulk measures of assembly *in vitro*. We find that methods based on derivative-free optimization (DFO) yields significant improvements in handling the special challenges of fitting self-assembly simulations relative to more traditional gradient-based methods. We use simulated data to examine how fit quality varies by data source, finding notable improvement for time-resolved mass spectrometry (MS) over static light scattering (SLS).**

***Keywords* — Self-Assembly, Stochastic Simulation, Data Fitting, Derivative Free Optimization, Virus Capsid.**

I. Motivation

VIRUS capsids have become key models for macromolecular assembly [1]. Unfortunately, we lack experimental methods to directly monitor fine-scale pathways of large, symmetric self-assembly systems like capsids. Simulation methods have therefore been proven invaluable for exploring possible pathways and understanding how they contribute to such robust and efficient assembly. Purely theoretical models, however, can only estimate ranges of possible pathways, not make conclusions about specific viruses, because their behavior is highly dependent on kinetic rate parameters that are not directly measureable.

We have previously addressed these challenges by indirectly learning quantitative assembly models of specific capsids by fitting rule-based stochastic simulations [2] to experimental SLS measurements of bulk capsid assembly *in vitro* [3,4]. This approach made it possible to predict possible fine-scale assembly pathways of specific real capsids. It also faced substantial challenges, though, due to computationally costly simulations, stochastic variation between trajectories, noisy experimental data, and degeneracy of possible solutions. We have sought to address these problems with a paired strategy of seeking improved data-fitting algorithms and exploring alternative experimentally feasible data sources better suited for unambiguously learning assembly models from bulk assembly assays.

Acknowledgements: This work was supported by NIH grant 1R01AI076318.
[1]Joint Carnegie Mellon – University of Pittsburgh Ph.D Program in Computational Biology and [2]Department of Biological Sciences and [3]Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University.

II. Modeling and Methods

We develop rule models for viral capsid as in our prior work and simulate them by a spatial stochastic simulation algorithm. We then fit rate parameters to minimize root mean square deviation (RMSD) between data points generated by simulations and objective datasets. We have specifically explored the use of DFO [5] methods, using two examples of different DFO paradigms, multi-coordinate search (MCS) [6] and stable noisy branch and fit (SNOBFIT) [7], in addition to the heuristic gradient-based method used in our prior work. We have further examined alternative experimental data sources, particularly time-resolved MS, relative to the SLS used in most studies of capsid assembly kinetics to date. We examine quality of fit for two real capsid systems, HPV and HBV, and an artificial tetrahedral test system, using a mixture of real and synthetic SLS and MS data. We evaluate methods by their ability to minimize RMSD on real and synthetic data and to recover true parameters for synthetic data sets.

III. Discussions

Our results suggest DFO algorithms can yield substantial improvements over the fitting methods typically used for simpler reaction systems. Likewise, newer technologies for monitoring bulk assembly can lead to more precise parameter fits thus better reconstruction of assembly pathways. Productively combining state-of-the-art data fitting algorithms and experimental biotechnology can be expected to yield great advances in our ability to reveal detailed mechanisms of even highly complex self-assembly processes.

References

[1] Whitesides GM, Grzybowski B (2002) Self-assembly at all scales. *Science* **295**, 2418-2421.

[2] Zhang T, Rohlfs R, Schwartz R (2005) Implementation of a discrete event simulator for biological self-assembly systems. *Proc. 37th Winter Simulation Conf.* pp. 2223-2231

[3] Kumar MS, Schwartz R (2010) A parameter estimation technique for stochastic self-assembly systems and its application to human papillomavirus self-assembly. Phys. Biol. **7**, 45005-45016.

[4] Xie, L, Smith, G.R., Feng, X., Schwartz, R. (2012) Surveying capsid assembly pathways through simulation-based data fitting. *Biophys. J.* **103**, 1545-1554.

[5] Rios, L. M. and N. V. Sahinidis, Derivative-free optimization: A review of algorithms and comparison of software implementations, Journal of Global Optimization, **56**, 1247-1293, 2013.

[6] Neumaier, A. MCS: Global Optimization by Multilevel Coordinate Search. http://www.mat.univie.ac.at/~neum/software/mcs/.

[7] Huyer, W., and A. Neumaier. 2008. SNOBFIT – Stable noisy optimization by branch and fit. *ACM Transactions on Mathematical Software* 35:1–25.